Wiki3DRank: un modelo para medir la relevancia de objetos de conocimiento mediante datos cuantitativos de Wikidata y Wikipedia

Wiki3DRank: A model for measuring the relevance of knowledge objects using quantitative data from Wikidata and Wikipedia

Juan Antonio PASTOR-SÁNCHEZ (1), Tomás SAORÍN (1), María-José BAÑOS MORENO (2)

(1) Department of Information Studies, University of Murcia. (2) Odilo / Department of Information Studies, University of Murcia. {pastor|tsp|mbm41963}@um.es

Resumen

Se presenta el modelo Wiki3DRank, que combina datos cuantitativos extraídos en tiempo real de Wikidata y Wikipedia para obtener un ranking de objetos de conocimiento a través de un valor cuantitativo que mida la relevancia de un objeto frente a otros en un determinado dominio. El modelo se basa en la distribución de los objetos de conocimiento en un espacio vectorial cuyas componentes se basan en tres variables principales: número de declaraciones en Wikidata sobre un ítem, número de artículos en las diferentes ediciones de Wikipedia y extensión en número de palabras de dichos artículos. Estas variables se asocian al nivel de descripción de los ítems de Wikidata, la difusión de los objetos de conocimiento asociados a los mismos en las ediciones de Wikipedia de diferentes idiomas y el grado de elaboración editorial de los correspondientes artículos de Wikipedia. Para demostrar la viabilidad del modelo se analizan una serie de casos de uso sobre diversos dominios: libros, películas, catedrales, terremotos, ríos y elementos químicos. A partir de los resultados obtenidos es posible concluir que Wiki3DRank es una herramienta que permite medir la relevancia de objetos de conocimientos en el contexto de un dominio de conocimiento. Se muestra el funcionamiento de una herramienta de código abierto que permite el cálculo en línea de Wiki3DRank. Los resultados obtenidos sugieren que el modelo propuesto puede aplicarse para diferentes contextos y dominios, que pueden introducirse elementos de ponderación y es posible extender el modelo mediante la introducción de nuevos componentes basados en otras características de los datos enciclopédicos de los objetos de conocimiento, al mismo tiempo que se mantiene el sistema de cálculo vectorial de base.

Palabras clave: Wiki3DRank. Rankings. Wikidata. Wikipedia. Conocimiento enciclopédico. Análisis de dominios. Objetos culturales.

Abstract

This research introduces the Wiki3DRank, a model combining real-time extracted quantitative data from Wikidata and Wikipedia to obtain a ranking of knowledge objects through a quantitative value that measures the relevance of one object compared to others in a specific domain. The model is based on the distribution of knowledge objects in a vector space, whose components are based on three main variables: the number of statements on Wikidata about an item, the number of articles in different Wikipedia editions, and the length in number of words of these articles. These variables are associated with the level of description of the Wikidata items, the dissemination of the referred knowledge objects in Wikipedia editions in different languages, and the degree of editorial elaboration of the corresponding Wikipedia articles. To demonstrate the viability of the model, a series of use cases across various domains are analysed: books, movies, cathedrals, earthquakes, rivers, and chemical elements. From the results obtained, it is possible to conclude that Wiki3DRank is a tool that allows measure the relevance of knowledge objects in the context of a knowledge domain. The operation of an open-source tool that enables the online calculation of Wiki3DRank is presented. The results suggest that the proposed model can be applied to different contexts and domains and that it`s ease to expand it by adding elements of weighting and extending the model with new components based on other characteristics of the encyclopaedic data of the knowledge objects, while the base vector calculation system is maintained.

Keywords: Wiki3DRank. Rankings. Wikidata. Wikipedia. Encyclopaedic knowledge. Domain analysis. Cultural objects.

1. Introducción

Este trabajo propone un método para calcular un ranking aplicable a los objetos de conocimiento registrados en Wikidata y Wikipedia. Valorar, reseñar y comentar son fenómenos sociales en sí mismos que forman parte de la discusión pública y que se manifiesta a través de medios clásicos

y actuales (Black, 2007). La cultura del ranking tiene una larga tradición, pero su forma contemporánea tiene un gran impacto a través de un progresivo proceso de cuantificación de las interacciones sociales. Ejemplos de ello son las listas de los más vendidos, los imprescindibles, los mejores tenistas del circuito, las ciudades con mejor calidad de vida, etc. La valoración y las puntuaciones conviven con las críticas, reseñas y estudios culturales. Todas ellas actúan como una especie de género periodístico que sirve como patrón para promocionar o acceder a la cultura en un sentido muy amplio. En el marco de las industrias editoriales, los contenidos audiovisuales y el entretenimiento, es muy acusada la importancia de aparecer en listas, ranking y selecciones. El objetivo de este artículo es explorar la generalización de las ideas plasmadas en un estudio anterior limitado a obras literarias (Pastor-Sánchez; Saorín; Baños-Moreno, 2023) situando su foco en lo que denominamos como objetos de conocimiento. Éstos se podrían definir preliminarmente como aquellas entidades de todo tipo que adquieren suficiente notoriedad para merecer un artículo en cualquiera de las ediciones de Wikipedia. El concepto "artículo de Wikipedia" implica una dificultad, al estar constituido en realidad por un número variable de artículos en muchos idiomas sobre un mismo elemento, hecho o concepto. Los artículos de las diferentes ediciones de Wikipedia se corresponden con un objeto de conocimiento que recoge tanto información como datos factuales relevantes para su explicación y comprensión. En el primer caso (información) se trataría de una fusión de toda la información escrita sobre un objeto en todos los idiomas de Wikipedia, y en el segundo (datos factuales) de los datos sintetizados en Wikidata para ese elemento individualizado.

Existe una amplia bibliografía para el análisis de la calidad de las enciclopedias y artículos, basadas en múltiples factores relacionados con el trabajo de edición colaborativa de artículos (Moás; Teixeira Lopes, 2023). A menudo se incorporan también aspectos de audiencia según el interés despertado por cada artículo. Por el contrario, son escasas las aproximaciones basadas en ratings y evaluaciones externas. Una de las metodologías más frecuentes es la del análisis de redes, que es una tendencia de largo alcance en la renovación de las investigaciones en humanidades y el campo cultural, fenómeno denominado "Network turn" (Ahnert; Ahnert; Coleman y Weingart, 2020). Sin embargo, dado que existen diversas Wikipedias para cada idioma, que constituyen un grafo propio, su análisis se presenta problemático desde el punto de vista de los objetos universales de conocimiento.

Wikipedia cubre todos los temas y sirve como cartografía del estado actual del conocimiento: es un mapa de conceptos continuamente enriquecido. Por lo tanto, ofrece un singular punto de entrada a la indagación sobre el ranking de cómo esos objetos son tratados a nivel informativo. Los estudios sobre cobertura temática en Wikipedia han girado sobre diversos campos, como el de la ciencia, las biografías, patrimonio cultural, cultura de masas o la actualidad social (Hill y Shaw, 2020; Reznik y Shatalov, 2016; Minguillón y otros, 2017). Wikipedia compite con un buen número de fuentes de información especializada en cada campo, como las bases de datos de cine, los repertorios de historia de la música o los catálogos de biblioteca. El discurso textual de Wikipedia se refuerza desde la puesta en marcha en 2012 de Wikidata. Se trata de una iniciativa que ofrece como infraestructura para almacenar datos estructurados derivados del contenido de los artículos de las diferentes ediciones de Wikipedia.

Debido al interés por el ranking de la sociedad actual es habitual encontrar listas basadas en propiedades intrínsecas cuantificables: longitud de los ríos, las ventas de libros, la fortuna de los millonarios, la población de las ciudades o el peso de los elementos químicos. También existen múltiples rankings sobre temas de interés social tales como libros, películas, políticos, deportistas o eventos. Esta clase de listas son actualizadas periódicamente por los medios de comunicación, e incluso constituyen una serie editorial global de los "Los 1001 … que hay que … antes de morir". Son menos frecuentes los ranking que intentan medir la importancia de elementos en temas tales como catedrales, batallas navales, yacimientos arqueológicos, mundiales de fútbol, papas de la Iglesia Católica o plantas aromáticas. Sin embargo, Wikidata puede entenderse como un sistema de objetos de conocimiento y un sistema de información cuyas características estructurales permiten acceder a sus datos. Por esto, el procesamiento de estos datos podría permitir generalizar un método de cálculo de ranking, que en este trabajo se denomina, Wiki3DRank, enfocado a la definición de indicadores sencillos de calcular y explicar.

Se han realizado numerosas propuestas para la evaluación automática de aspectos de calidad de los contenidos de Wikipedia basados en métodos cuantitativos. Estas propuestas constituyen por sí mismas un subcampo de estudio sobre Wikipedia (Nielsen, 2012). Algunos autores explotan las métricas del análisis de redes, otros usan las métricas propias disponibles para el contenido de los artículos: número de palabras, número de referencias, extensión, enlaces entrantes, etc., complementados con el estudio de la actividad

de los editores, reputación y redes de colaboración. De modo parecido sucede en Wikidata en relación con las investigaciones para establecer la calidad y completitud de los datos (Shenoy, 2022). Se trata de un campo que también genera investigación aplicada. El trabajo más conocido sobre ranking es el de Skiena y Ward (2013) en el que se comparan personajes históricos diferenciando entre celebrity (popularidad actual) y gravitas (popularidad consolidada). De forma similar, la base de datos Networked Pantheon desarrolla la aplicación de medidas de centralidad y similares en el grafo de de biografías resultante de Wikipedia (Beytía; Schobin, 2020).

Un proyecto de investigación relevante en el ámbito de este trabajo es WikiRank (Lewoniewski y otros, 2019). Actualmente mantiene un servicio en línea (www.wikirank.net) que permite obtener rankings multilingües de artículos. Este servicio es un ejemplo de cómo definir rankings de artículos segmentados por tipos de contenido usando indicadores agregados que denominan "popularity", "Authors' Interest" y "Citation Index". Se basa en el procesamiento periódico de dumps de Wikipedia, y permite acceder también a datos históricos de evolución del ranking. Es posible observar una versión de Wikipedia en concreto y también por categorías temáticas para realizar comparativas entre artículos dentro de agrupaciones temáticas obtenidas a partir de la exploración de categorías, clases de Wikidata y la ontología Dbpedia. Cada artículo recibe un valor de 1 a 100, basado en el análisis de las características más frecuentes usadas en los estudios sobre evaluación automática de calidad en Wikipedia: completeness, credibility, objectivity, readability, relevance, style, timeliness. Se toman los datos de longitud de los artículos, número de referencias, densidad de referencias, número de imágenes y número de secciones, que son obtenidos desde la herramienta Wikimedia XTools, y se obtiene una medida sintética a partir de valores normalizados para cada característica y adaptados a cada versión de Wikipedia. Se asigna la puntuación máxima (100) en aquellas características en las que un artículo supera la mediana del idioma correspondiente. Posteriormente se realiza una media aritmética de las características evaluadas y finalmente se modula teniendo en cuenta la existencia de plantillas de control de calidad en el artículo, y de esta manera se mide la calidad. WikiRank también se mide la popularidad, con medidas sintéticas sobre visualizaciones de página y número de editores. Cada bloque temático muestra los artículos más populares, su cobertura en los idiomas analizados y cuál es la versión con mayor calidad. De cada artículo es posible obtener una visión de cómo evoluciona su popularidad a lo largo del tiempo, de forma global y dentro de cada Wikipedia. También es posible obtener indicadores del número de enlaces de cada artículo dentro de su propia Wikipedia, y un valor global acumulado (Citation Index).

Existe cierta similitud entre las denominaciones WikiRank y el modelo Wiki3DRank propuesto en este trabajo. Sin embargo, es necesario destacar que Wiki3DRank se centra en el uso de una medida sintética de los objetos de conocimiento enciclopédico. Esta medida se basa tanto en el análisis de características del correspondiente ítem de Wikidata como de sus artículos equivalentes en Wikipedia. Además, se adopta un modelo conceptual basado en la representación de los objetos dentro de un espacio vectorial cuya dimensionalidad, como se expone en el apartado de discusión, puede adaptarse a diferentes escenarios de análisis.

Por lo tanto, este trabajo define una serie de objetivos y una metodología de trabajo para determinar los datos necesarios y el modo en el que se deben obtener, procesar y utilizar para obtener una medida, Wiki3DRank, que permita identificar y ponderar objetos de conocimiento multidominio derivados del uso combinado de Wikidata y Wikipedia.

2. Objetos de conocimiento: de los artículos de Wikipedia a los elementos de Wikidata

Las enciclopedias tienen una larga tradición en respuesta a la necesidad de delimitar el conocimiento básico disponible en un momento dado y presentarlo en un formato accesible (Brown, 2011): compacto, orientado a la explicación precisa de varios aspectos de un concepto: surgimiento, evolución, aplicaciones, conexiones. Es interesante contemplar Wikipedia desde dos puntos de vista importantes para nuestro propósito, su cobertura temática y la extensión de contenido.

Desde el punto de vista de su cobertura, Wikipedia ha alcanzado una amplitud de temas tratados nunca vista con anterioridad. También destaca por su capacidad de respuesta rápida para incorporar información sobre nuevos acontecimientos. Su crecimiento es continuo porque la propia realidad genera nuevos datos y entidades merecedoras de atención. Además, la combinación de su formato digital y su política editorial distribuida, ha facilitado el "inclusionismo", que amplía enormemente el rango de que lo se admite como relevancia o notabilidad enciclopédica (McDowell; Vetter, 2022, pp. 46-70). En un contexto digital y con una gran masa de editores, se pueden asumir artículos sobre muchos más temas. Al mismo tiempo, permite un alto nivel de granularidad, puesto que cada parte específica de un tema puede abordarse en un artículo propio.

Desde el punto de vista de la extensión del contenido, los artículos de Wikipedia manifiestan mayor parecido a los artículos de enciclopedias especializadas que a las genéricas. Esto se debe a que los artículos tienden a alcanzar una extensión considerable, se dividen en secciones, incluyen notas y están densamente conectados con otros conceptos de la enciclopedia. Aunque el ideal enciclopédico es el de presentar de forma suficiente un tema, la elasticidad de la página digital permite que los editores añadan información relevante para ofrecer un panorama amplio del asunto tratado, desde varios puntos de vista. La exigencia de verificabilidad hace que los artículos contengan además una bibliografía básica de orientación en cada tema, así como notas con referencias a publicaciones especializadas.

Muchos de los estudios sobre Wikipedia se realizan sobre la versión en inglés, dando por supuesto que, al ser la que contiene un mayor número de artículos, refleja casi todo el conocimiento global. Sin embargo, es necesario recordar que cada Wikipedia es un proyecto independiente, y que se detecta un amplísimo número de artículos sobre temas que no están presentes en inglés (Miquel-Ribé, 2019). Existe, desde luego, un importante grado de coincidencia de artículos de diferentes ediciones de Wikipedia sobre un mismo asunto, tema o concepto. Pero también son importante las diferencias por la falta de cobertura entre idiomas. La puesta en marcha de Wikidata como base de datos factual que conecta entre sí todas las Wikipedias, pone de manifiesto la existencia de un mapa de conceptos global resultado de la suma de todas las enciclopedias, independiente del idioma en el que se origina.

Aquí entra en juego un tercer aspecto, además de los ya mencionados relacionados con la convergencia y extensión de temas, que es el artefacto Wikidata-Wikipedia como sistema de organización del conocimiento. La comprensión de este sistema requiere cada vez más un uso más ágil de las técnicas de análisis de dominio (Smiraglia, 2015). Una enciclopedia, especialmente Wikipedia, es a la vez un vocabulario científico-cultural y un inventario de nombres de autoridad para individuos y grupos, a la vez que una enorme atención dedicada al registro de eventos sociales. La categorización nativa en Wikipedia es una mezcla irregular de navegación, descripción y agrupación, pero no es una taxonomía muy adecuada para explorar un dominio de conocimiento. Sin embargo, aunque el sistema de clasificación en Wikidata se adhiere más estrechamente a los criterios estándar para una taxonomía correcta, exhibe muchas inconsistencias en su estructura jerárquica y asignaciones de clases.

El concepto de “objeto de conocimiento” considerado en este artículo está claramente vinculado al concepto de elemento (ítem) en Wikidata. Los artículos surgen en un idioma determinado, y se les asigna en Wikidata un identificador único de elemento que será usado para vincular los artículos que vayan surgiendo en otros idiomas con su ítem correspondiente. Un ítem, por ejemplo Q63167656 del artículo sobre el incendio de la catedral de Notre-Dame en 2019, conecta entre sí a los artículos en 58 Wikipedias, construyendo así una entidad individualizada para un concepto relevante, en este caso, del tipo acontecimiento.

Wikidata es un grafo de conocimiento que utiliza su propio modelo de datos compatible con RDF. Sus elementos principales son ítems que representan un objeto real, un concepto o un evento. Cada ítem tiene asociado un identificador único cuya designación comienza por la letra “Q”. Por ejemplo, el libro “Cien años de soledad” de Gabriel García Márquez es el elemento Q178869, aunque está vinculado a 74 artículos en diferentes Wikipedias (español, japonés, italiano, ruso, etc.). A su vez, cada ítem se describe mediante propiedades cuyas designaciones comienzan por la letra “P”. Las propiedades definen relaciones entre elementos o se refieren a valores literales (cadenas, números, fechas). Por ejemplo, del libro mencionando se declara que tiene como autor (P51) al elemento Q5878 (el escritor García Márquez) y que su fecha de publicación (P577) es 1967. Wikidata no tiene clases definidas explícitamente diferenciadas del resto de los elementos. En cambio, algunos elementos desempeñan tal papel de clase al enmarcarse en una taxonomía de clases y subclases conectadas a través de la propiedad P279 (subclase de). La pertenencia de los ítems a las clases se realiza mediante la propiedad P31 (instancia de). Es decir, Wikidata puede entenderse hasta cierto punto como una “ontología colaborativa” que contiene tanto datos primarios como el esquema utilizado para formalizar la organización del conocimiento (Piscopo y Simperl, 2018). Dentro de cada ítem existe una sección denominada “Identificadores”, que definen conexiones con registros y bases de datos externas de todo tipo, como, por ejemplo, con el sistema internacional de control de autoridades VIAF (Bianchini y Sardo, 2022). No todos los ítems de Wikidata tienen un artículo en alguna edición de Wikipedia, puesto que pueden ser creados como datos y no suscitar interés suficiente para requerir un artículo explicativo. Aunque existen elementos sin artículo, este trabajo se centra solo en aquellos que sí lo tienen, estableciendo por tanto un límite en objetos de conocimiento que han conseguido relevancia enciclopédica.

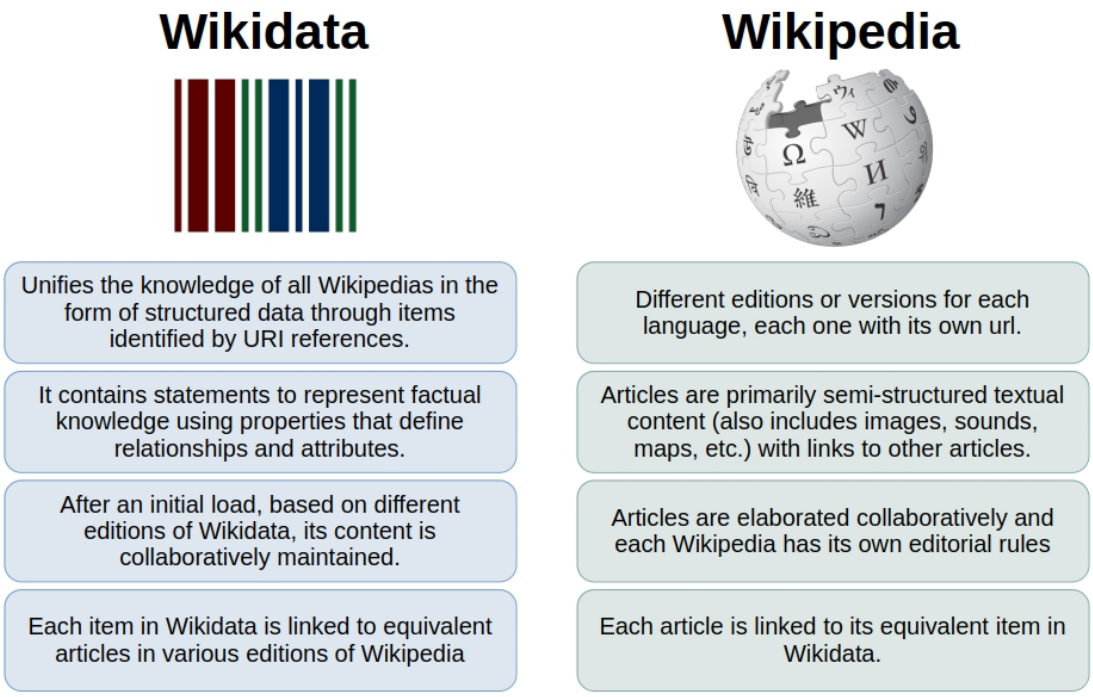

Figura 1. Resumen de las principales diferencias y conexiones entre Wikidata y las ediciones de Wikipedia

Un registro descriptivo de Wikidata puede incluir los siguientes tipos de información:

- *Etiquetas, alias y descripciones:* textos en diversos idiomas que permiten nombrar los ítems, alias (sinónimos) y descripciones con definiciones abreviadas. No todos los ítems tienen etiquetas y descripciones en los mismos idiomas.
- *Propiedades o atributos:* es el descriptor para un valor (ya sea literal u otro ítem) de una afirmación o declaración. Poseen un identificador que comienza por P (p.e. P23). Algunas propiedades permiten definir relaciones de instancia-clase (P31 instancia de) o establecer taxonomías con esquemas subclase-clase (P279 subclase de).
- *Declaraciones (statements):* datos sobre un determinado ítem. Formado por una afirmación con sus correspondientes cualificadores referencias y rangos.
- *Afirmaciones (claims):* datos para una propiedad concreta sobre un determinado elemento, que generalmente tienen la forma de vínculo con otra entidad de Wikidata. Formados por pares propiedad-valor.

Las declaraciones, a su vez, pueden estar especificadas mediante:

- *Calificadores (qualifiers):* afirmación que dice algo sobre una afirmación específica para matizarla o detallarla (reificación). También están formados por pares propiedad-valor.
- *Referencias (references):* describen el origen de una afirmación y pueden ser un enlace externo u otro ítem de Wikidata.
- *Rango (rankings):* indicador que permite identificar la declaración más relevante frente a otras cuando existen varias sobre una misma propiedad.

La idea que subyace a la propuesta de este trabajo es que Wikipedia puede considerarse al mismo tiempo discurso y datos. Los artículos de Wikipedia distribuídos en cada una de las ediciones de idiomas diferentes con una visión diferentes del conocimiento. Los datos, multilingües y agrupados en Wikidata que combina a todas las ediciones de Wikipedia y da forma a un inventario global de hechos o conceptos.

3. Metodología de cálculo de Wiki3DRank

El objetivo de este trabajo es proponer un método de cálculo de ranking de objetos de conocimiento de Wikipedia y Wikidata denominado *Wiki3DRank*. Esta métrica utiliza los contenidos de Wikipedia o Wikidata como proxy para medir algo relativo a la atención global sobre el conocimiento de un objeto externo; por tanto, no trata de medir la calidad de los artículos de Wikipedia o las descripciones de Wikidata.

Su funcionamiento se basa en gran medida en un trabajo anterior (Pastor-Sánchez; Saorín; Baños-Moreno, 2023), si bien supone un refinamiento desde el punto de vista de la definición conceptual del mismo. En el mencionado trabajo se trabajó en la definición de un canon literario a partir del número de declaraciones de un ítem de Wikidata, el número de ediciones de Wikipedia en las que dicho ítem tenía un artículo y la suma de palabras de dichos artículos, y se complementó con cálculos de clustering para delimitar subconjuntos coherentes. Para el cálculo del *Wiki3DRank* se definen, porta tanto, tres indicadores fundamentales, a saber:

- N_{Props} *:* Número de propiedades utilizadas para describir un ítem de Wikidata, exceptuando las propiedades utilizadas en la sección de identificadores. Este indicador refleja la profundidad y extensión en el proceso de descripción de un ítem.
- N_{Wikis} *:* Número de ediciones de Wikipedia en las que un ítem de Wikidata tiene un artículo correspondiente. Se trata de un indicador que representa el alcance a nivel global que tiene un determinado ítem en el contexto de diversos idiomas.
- N_{Words} *:* Se calcula a partir de la suma del número de palabras del contenido de todos los artículos de las diferentes ediciones de Wikipedia vinculados al ítem de Wikidata. Se trata de un indicador que mide el volumen de trabajo editorial realizado en el proceso de redacción de los artículos.

Es necesario tener en cuenta que los intervalos numéricos en los que oscilan estos indicadores es muy dispar. Las magnitudes medidas por N_{Props}, N_{Wikis} y N_{Words} son de naturaleza diferente. Por ejemplo: los valores que puede alcanzar N_{Words} al sumar todas las palabras de los artículos equivalentes de un ítem en todas las ediciones de Wikipedia es considerablemente mayor que el que pueden alcanzar N_{Props} y N_{Wikis}. Por otro lado, la naturaleza colaborativa tanto de Wikipedia como de Wikidata implica que la comunidad global de editores preste más atención a un conjunto relativamente reducido ítems y sus correspondientes artículos, mientras que otros tienen un desarrollo mucho más reducido.

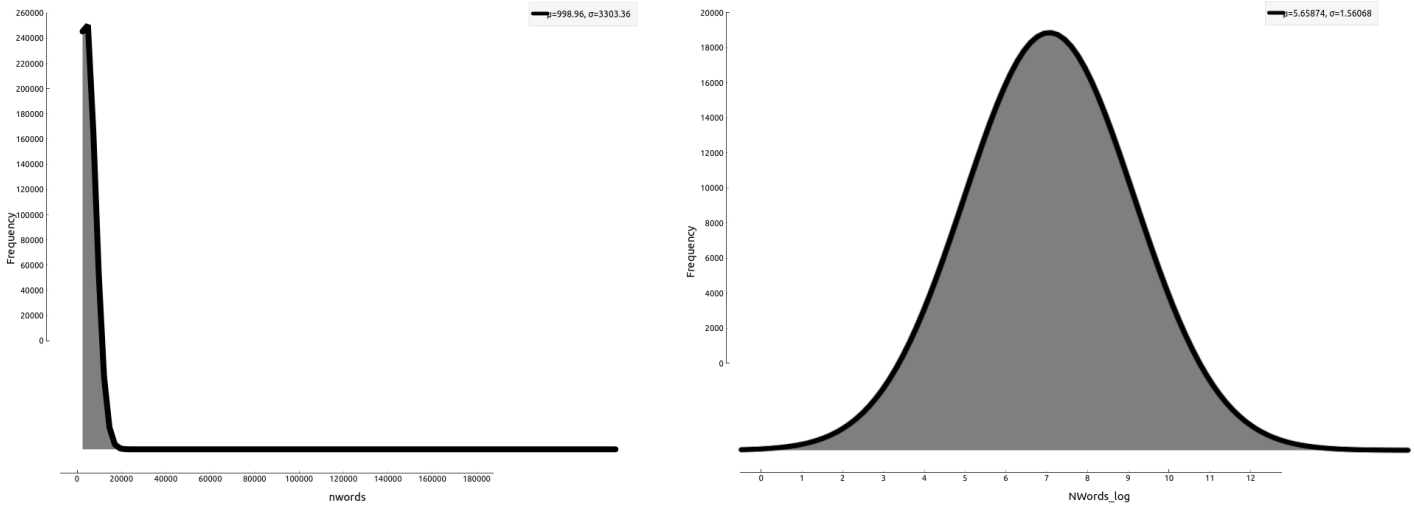

Figura 2. Distribución original de N_{Words} de un dataset de ítems de Wikidata sobre películas (izquierda) y distribución tras aplicar la transformación logarítmica log(1+N_{Words})

Por lo tanto, a priori, las distribuciones de los tres indicadores tienen una fuerte asimetría positiva característica del fenómeno de "larga cola" ampliamente estudiado en muchos procesos sociales, especialmente en comunidades y plataformas digitales (Anderson, 2014). Esto se debe a que existe una gran cantidad de ítems con unos valores bajos o intermedios para los tres indicadores y una pequeña cantidad de ítems que tienen valores altos para los mismos. Por ambos motivos se ha optado por aplicar una transformación logarítmica para combinar los tres indicadores. Esto permite normalizar la distribución de los tres indicadores y trabajar con indicadores cuyo intervalo original de valores es muy diferente y comparar diferentes muestras de datos. Además, se eligió la transformación logarítmica (en lugar de alternativas como Z-score, Min-Max o Robust Scaling) porque esta normalización puede realizarse de forma independiente en cada elemento de Wikidata, sin depender de los valores alcanzados por otros elementos de un

conjunto de datos. Esta característica permite el cálculo aislado de Wiki3DRank en tiempo real de forma rápida y sencilla.

La Figura 2, en la página anterior, muestra la distribución original de *NWords* de un dataset de ítems de películas de Wikidata y la obtenida tras la transformación logarítmica.

Considerando la aplicación de la mencionada transformación logarítmica se elaboró una primera propuesta para calcular *Wiki3DRank* para cada ítem como la agregación de tres componentes como muestra la siguiente ecuación:

$$a = log\ (1 + N_{Wikis})$$

$$b = log\ (1 + N_{Props})$$

$$c = log\ (1 + N_{Words})$$

$$Wiki3DRank = a + b + c$$

Los resultados obtenidos para la realización del trabajo previo sobre el canon literario validaron el método de cálculo del ranking. En este sentido se demostró que el uso combinado de los tres indicadores permitía obtener resultados más coherentes y precisos que el uso aislado de uno de ellos los indicadores o reduciendo la dimensionalidad mediante el análisis de componentes principales.

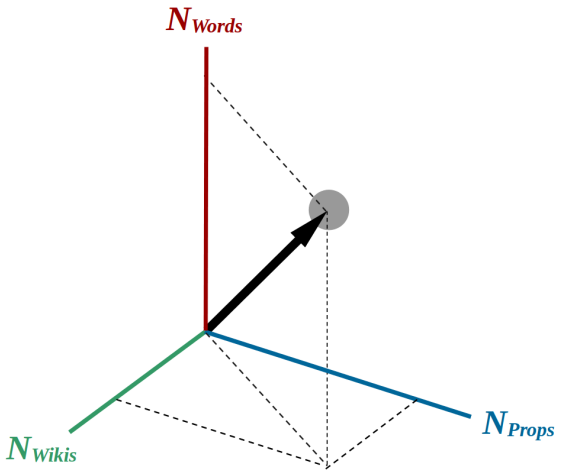

Figura 3. Representación de un ítem de Wikidata como un vector

Siguiendo esta aproximación, este trabajo adopta un enfoque más general, proponiendo un modelo en el cada ítem se representa como un vector, cuyas componentes serían, en un principio, los tres indicadores mencionados. En consecuencia, *Wiki3DRank* se podría calcular como el módulo de dicho vector (Figura 3).

Por lo tanto, en este trabajo se propone utilizar el módulo del vector para el cálculo de Wiki3DRank. De este modo se tendría que para cada ítem lo siguiente:

$$a = log\ (1 + N_{Wikis})$$

$$b = log\ (1 + N_{Props})$$

$$c = log\ (1 + N_{Words})$$

$$Wiki3DRank = \sqrt{a^2 + b^2 + c^2}$$

Este nuevo enfoque de *Wiki3DRank* permite representar los ítems de Wikidata dentro de un espacio vectorial. También ofrece un mecanismo de generalización que, como se muestra en la sección de discusión, permite incorporar nuevas componentes a los vectores de los ítems.

Un aspecto a destacar es que el método propuesto permite evaluar los ítems de forma independiente, uno a uno, sin necesidad de volver a procesar todo el conjunto de datos. En otras palabras, el cálculo de un Wiki3DRank para un elemento no depende en ningún momento del obtenido para otros elementos. Esto permite obtener un valor que puede compararse con el obtenido para otros ítems y establecer así una clasificación según sea necesario.

Para implementar el cálculo de Wiki3DRank es necesario acceder a los datos en los que se basa, *NWikis*, *NProps* y *NWords*, para cada ítem concreto que se necesite evaluar. Para todos ellos existen fuentes de datos en línea suficientes. En el caso de *NWikis* y *NProps* puede utilizarse Wikidata Query Service (WDQS, https://query.wikidata.org) para obtener aquellos datos correspondientes de cada ítem, a través de consultas SPARQL. En el caso de NWords, el acceso será a través de la API de Xtools (https://xtools.wmcloud.org/api). En el caso de *NProps*, no se han tenido en cuenta las declaraciones sobre propiedades que son meros identificadores, según el modelo de datos de Wikidata. Se detallan otros datos complementarios que se han incorporado al script para enriquecer la aplicación de consulta *Wiki3DRank*, y que se comentan en la sección de discusión.

4. Resultados

Este trabajo presenta dos tipos de resultados. En primer lugar, se realiza un análisis de casos de la aplicación multidominio de *Wiki3DRank* evaluando los datos al aplicarlo sobre obras creativas, objetos científicos, realidad geográfica, acontecimientos y monumentos arquitectónicos. En segundo lugar, se presenta una aplicación web que permite el cálculo de *Wiki3DRank* en línea y en tiempo real. Los conjuntos datos y los

scripts utilizados para la generación y procesamiento de los mismos están disponibles en el repositorio Zenodo, el código fuente de la aplicación *Wiki3dRank Calculation* está disponible en GitHub y el aplicativo web alojado en los servidores web de la Universidad de Murcia, en las direcciones que se indican al final del trabajo.

4.1. Análisis de casos de uso

Esta sección presenta diferentes casos de uso de rankings para objetos de conocimiento, con el fin de tener una primera aproximación a su uso en contextos prácticos. Para captar mejor el comportamiento de la medida propuesta, se han seleccionado los siguientes casos específicos de dominios con diferentes características: Obras literarias, películas, elementos químicos, ríos, terremotos y catedrales.

Para el propósito exploratorio de esta parte de la investigación, consideraremos el concepto de dominio sin establecer una definición formal y buscando un enfoque generalista. Las obras literarias y las películas se entienden en el contexto de las obras creativas, como realizaciones culturales que se distribuyen de forma masiva y que contienen un contenido único vinculado a la autoría y la originalidad. Los elementos químicos constituyen un conocimiento universal perfectamente delimitado en las ciencias básicas, estable y limitado a unos pocos ítems reunidos en la tabla periódica. Los ríos son un tipo de objeto de conocimiento presente en todo el planeta, abundante y estudiado desde la geografía física y otras disciplinas. Los terremotos son acontecimientos imprevistos de gran impacto social, con una larguísima historia y diferencias significativas en su intensidad y consecuencias, mientras que las catedrales son elementos materiales característicos de la cultura cristiana, y constituyen un tipo de monumento muy reconocible y objeto de atención no solo desde la historia del arte, sino también desde otros campos como el turismo o la interpretación del patrimonio. En todos ellos se combinan de forma diferente propiedades o atributos como el factor tiempo, objetividad, interpretación, idioma, impacto mediático, cambio, materialidad, ámbito desde el que se estudia o universalidad. Conviene señalar, no obstante, que se han elegido casos en los que, en principio, es sencillo delimitar qué entra dentro de esa categoría. No obstante, hay que considerar que cualquier actividad de recolección requiere en su fase inicial delimitar operativamente qué entra y no entra dentro de una categoría. Los objetos no son puros, sino que se accede a ellos a través de un punto de vista; un ejemplo de ello sería afluentes de ríos, partes de un conjunto arquitectónico u obras en serie.

Para la selección de los items se ha usado la potencia de consulta que ofrece Wikidata, a través de la tipificación “Instancia de” (P31). Aunque la categorización de Wikidata adolece de problemas de nivel de detalle en su asignación, así como de falta de rigor en la definición de clases y subclases (Piscopo, 2019), la consulta directa por clases comunes permite obtener conjuntos significativos de elementos y precisos. Para cada caso de uso se ha recuperado el item de dominio más frecuente, de forma directa, sin tener que recurrir a complejas consultas recursivas. Pueden quedar fuera algunos objetos relevantes, pero para los fines ilustrativos de este trabajo es suficiente. Se obtienen los siguientes resultados globales:

	Obras literarias	*Películas*	*Elementos químicos*	*Ríos*	*Terremotos*	*Catedrales*
Item de dominio	Q7725634	Q11424	Q11344	Q4022	Q7944	Q56242215
Nº items	118497	267177	166	411443	2217	855
Declaraciones	1070183 (*)	7760860	13384	2973093	16889	19153
Declaraciones no ID	794293 (*)	4156620	5230	2154220	12467	13884
Artículos Wikipedias	249263	1013165	18116	723662	9237	8344
Correl. N_{Wikis}-N_{Props}	0.534 0.383	0.769 0.693	0.802 0.935	0.745 0.531	0.581 0.373	0.827 0.809
Correl. N_{Wikis}-N_{Words}	0.854 0.545	0.823 0.786	0.776 0.949	0.766 0.645	0.876 0.744	0.891 0.891
Correl. N_{Props}-N_{Words}	0.496 0.327	0.630 0.661	0.820 0.929	0.567 0.432	0.495 0.293	0.843 0.843

Tabla I. Resumen de los dominios de ítems de los casos de uso (Datos: enero 2024)

En La Tabla I se señala con (*) que durante la fase de obtención de datos se detectó una anomalía en el ítem Q213019 correspondiente a la obra literaria “La Guerra de los Mundos” de George Orwell. Dicha anomalía consiste en la reciente introducción de 6.400 declaraciones de

traducciones o ediciones de dicha obra. Las declaraciones fueron creadas por un bot entre el 13 y el 17 de enero de 2023. Para este trabajo se ha procedido a no considerar dichas declaraciones debido a que supone un valor extremo que altera significativamente los datos estadísticos. La Tabla incluye los datos descontando las declaraciones mencionadas en el dominio de Obras literarias.

Los 20 primeros elementos ordenados por ranking en cada dominio, serían los siguientes:

Elementos químicos			*Ríos*		
Item	*Label*	*Wiki3DRank*	*Item*	*Label*	*Wiki3DRank*
Q897	gold	14.08077	Q1653	Danube	14.33553
Q677	iron	13.98320	Q584	Rhine	13.80449
Q753	copper	13.98121	Q3783	Amazon	13.53747
Q629	oxygen	13.97657	Q626	Volga	13.53228
Q556	hydrogen	13.83113	Q3392	Nile	13.52990
Q568	lithium	13.82849	Q5089	Ganges	13.44008
Q663	aluminium	13.73210	Q1644	Elbe	13.24763
Q623	carbon	13.70546	Q5413	Yangtze	13.21684
Q560	helium	13.69960	Q7348	Indus River	13.12271
Q1090	silver	13.65123	Q2251	Columbia River	13.07417
Q627	nitrogen	13.50882	Q7355	Yellow River	13.06024
Q925	mercury	13.41739	Q5419	Missouri River	13.04587
Q708	lead	13.41456	Q3503	Congo	12.97481
Q1098	uranium	13.41390	Q973	Ob	12.88339
Q758	zinc	13.40793	Q3542	Niger River	12.79182
Q871	arsenic	13.29793	Q40855	Dnieper	12.71570
Q716	titanium	13.28139	Q19686	River Thames	12.68431
Q682	sulfur	13.25536	Q1265	Colorado River	12.63784
Q674	phosphorus	13.20026	Q41986	Meuse	12.56699
Q725	chromium	13.17737	Q78707	Yenisey	12.55018

Tabla II. Lista de los veinte primeros resultados ordenados por Wiki3DRank para elementos químicos y ríos (Datos: enero 2024)

En la Tabla II puede comprobarse cómo los elementos químicos, al ser un concepto de ciencia básica y un conjunto cerrado de items, se ajusta menos al rankeo, con diferencias muy estrechas, mientras que en los ríos se sugiere una estrecha relación entre su extensión, y por tanto su impacto en el territorio y su posición en el ranking.

Los datos respectos a terremotos y catedrales se incluyen en la Tabla III.

Terremotos			*Catedrales*		
Item	*Label*	*Wiki3DRank*	*Item*	*Label*	*Wiki3DRank*
Q43777	2010 Haiti earthquake	12.25761	Q2981	Notre-Dame de Paris	13.57579
Q122351413	2023 Marrakesh-Safi earthquake	11.80661	Q5943	St. Stephen's Cathedral	12.90083
Q151835	2010 Chile earthquake	11.69395	Q4176	Cologne Cathedral	12.87427
Q19830062	April 2015 Nepal earthquake	11.66956	Q205136	Cathedral of Santiago de Compostela	12.86502
Q191055	1755 Lisbon earthquake	11.63256	Q180274	Notre-Dame de Chartres	12.62891
Q211386	1906 San Francisco earthquake	11.40086	Q106934	Notre-Dame d'Amiens	12.55677
Q1798567	1985 Mexico City earthquake	11.35079	Q18068	Milan Cathedral	12.51069
Q152033	2008 Sichuan earthquake	11.30518	Q1123180	Toledo Cathedral	12.46980
Q212618	1960 Valdivia earthquake	11.24643	Q231606	Catedral de Sevilla	12.46332
Q207918	2009 L'Aquila earthquake	11.22820	Q33200	Mosque-Cathedral of Cordoba	12.40758

Q214866	Great Hanshin earthquake	11.12279	Q84090	Archbasilica of St. John Lateran	12.39620
Q1348910	1908 Messina earthquake	11.08844	Q610961	Mexico City Metropolitan Cathedral	12.38517
Q56768333	2018 Sulawesi earthquake and tsunami	11.05736	Q5949	St. Vitus Cathedral	12.28075
Q112666390	June 2022 Afghanistan earthquake	10.97984	Q389210	Pamplona Cathedral	12.19373
Q191293	1556 Shaanxi earthquake	10.94832	Q184407	Basilica of Saint-Denis	12.17310
Q104535090	2020 Petrinja earthquake	10.89959	Q17155	Cathedral of the Holy Cross and Saint Eulalia	12.14300
Q151850	February 2011 Christchurch earthquake	10.89246	Q745460	Cathedral of Our Lady of Strasbourg	12.14084
Q462195	1976 Tangshan earthquake	10.85084	Q744420	Burgos Cathedral	12.09830
Q115322003	2022 Cianjur earthquake	10.82147	Q206823	Reims Cathedral	12.05725
Q189079	2011 Van earthquake	10.77991	Q22720	Speyer Cathedral	11.93270

Tabla III. Lista de los veinte primeros resultados ordenados por Wiki3DRank para terremotos y catedrales (Datos: enero 2024)

Se observa que en los terremotos recogidos se percibe su tratamiento como fenómeno con significado cultural frente a meros aspectos geofísicos. Los terremotos contemporáneos tienen mayor cobertura en los medios, pero las grandes catástrofes históricas mantienen su relevancia. Las catedrales destacan, lógicamente, por su valor monumental y turístico, aquellas de un periodo concreto de la historia cristiana europea y de ultramar.

Por su parte, la Tabla IV muestra los resultados para Obras literarias y Películas.

Obras literarias			*Películas*		
Item	*Label*	*Wiki3DRank*	*Item*	*Label*	*Wiki3DRank*
Q428	Qur'an	14.61518	Q44578	Titanic	14.12210
Q9184	Book of Genesis	14.02116	Q24871	Avatar	13.79666
Q8275	Iliad	13.88599	Q17738	Star Wars: Episode IV – A New Hope	13.76991
Q35160	Odyssey	13.61286	Q163872	The Dark Knight	13.73499
Q480	Don Quixote	13.51878	Q47703	The Godfather	13.70764
Q43361	Harry Potter and the Philosopher's Stone	13.46623	Q104123	Pulp Fiction	13.66943
Q74287	The Hobbit	13.44176	Q23781155	Avengers: Endgame	13.65912
Q6511	Ulysses	13.40296	Q102438	Harry Potter and the Philosopher's Stone	13.61317
Q8272	Epic of Gilgamesh	13.30937	Q2875	Gone with the Wind	13.61120
Q8258	One Thousand and One Nights	13.29926	Q134430	Snow White and the Seven Dwarfs	13.56524
Q92640	Alice's Adventures in Wonderland	13.29783	Q23780914	Avengers: Infinity War	13.46630
Q208460	Nineteen Eighty-Four	13.29518	Q182218	The Avengers	13.44259
Q161531	War and Peace	13.27030	Q91540	Back to the Future	13.42724
Q8279	Shahnameh	13.23405	Q103474	2001: A Space Odyssey	13.42255
Q60220	Aeneid	13.21705	Q18407657	Captain America: Civil War	13.41430
Q150827	Frankenstein; or, The Modern Prometheus	13.14119	Q23780734	Black Panther	13.36900
Q19786	Old Testament	13.13063	Q14171368	Avengers: Age of Ultron	13.36220
Q165318	Crime and Punishment	13.11137	Q134773	Forrest Gump	13.35778
Q178869	One Hundred Years of Solitude	13.10061	Q132689	Casablanca	13.33334
Q46758	Harry Potter and the Deathly Hallows	13.09859	Q483941	Schindler's List	13.32423

Tabla IV. Lista de los veinte primeros resultados ordenados por Wiki3DRank para obras literarias y películas (Datos: enero 2024)

En la Tabla IV se observa en la literatura cierto equilibrio entre épocas, culturas de origen y géneros. Resulta interesante la presencia de obras mitológicas y textos religiosos. En el caso

del cine, con una historia de poco más de un siglo, se observa un claro predominio del cine americano. También se detecta un fenómeno de preponderancia de películas muy actuales y destaca la presencia entre los 20 primeros resultados de seis películas de la franquicia Marvel.

4.2. Web app for real-time Wiki3DRank Calculation

El modelo utilizado para la representación de Wiki3DRank permite la obtención de los datos para su cálculo en línea y en tiempo real. El cálculo de Wiki3DRank para un ítem no requiere el procesamiento de volcados masivos de datos de Wikipedia o de Wikidata. Es posible, mediante consultas a WDQS y Xtools, obtener los datos de un modo relativamente sencillo y rápido.

Como parte de los resultados de este trabajo, se ha publicado una aplicación lista para usar para el cálculo de Wiki3DRank. Esta aplicación ha sido desarrollada en Python (recuperación de datos) y PHP (cálculo de Wiki3DRank y visualización de resultados). El funcionamiento es sencillo: el usuario sólo tiene que introducir uno o varios códigos de artículos de Wikidata, y la aplicación se encarga de recuperar los datos, realizar los cálculos y mostrar los resultados (véase la Figura 4).

El funcionamiento es sencillo: el usuario sólo tiene que introducir uno o varios códigos de elementos de Wikidata, y la aplicación se encarga de recuperar los datos, realizar los cálculos y mostrar los resultados (véase la Figura 2). Una característica interesante es que permite seleccionar por separado los componentes de cálculo que se mostrarán y los que se utilizarán para el cálculo del Wiki3DRank. Dado nuestro propósito demostrativo, se ofrecen varias formas alternativas de calcular el Wiki3DRank, que se explicarán en la siguiente sección de este trabajo.

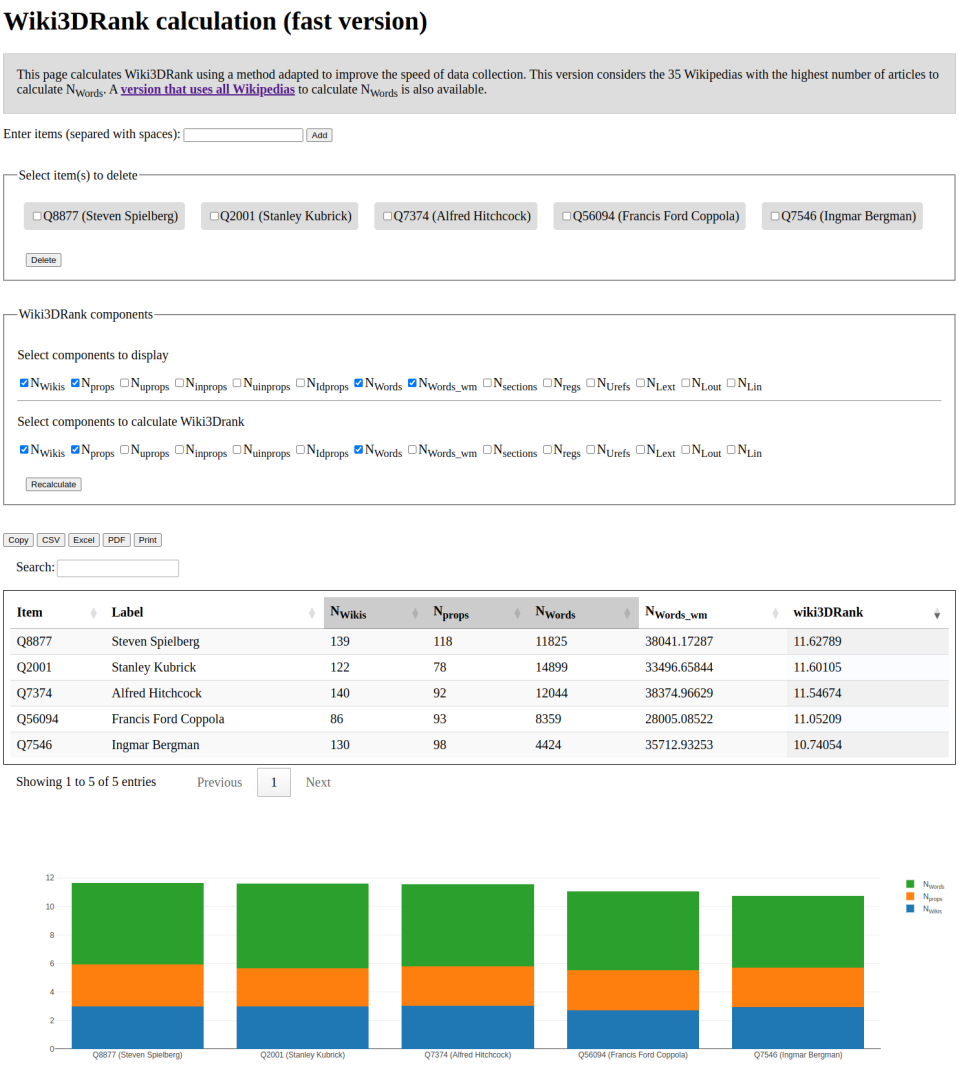

Figura 4. Ejemplo de uso de Wiki3DRank Calculator para los ítems de varios directores de cine (Datos: enero 2024, https://gicd.inf.um.es/wiki3drank)

No se pueden recuperar más de 20 elementos simultáneamente. Se muestran en una tabla detallada y en un gráfico de columnas apiladas, ambos exportables, y hay disponibles componentes adicionales y una versión de recuperación más rápida para realizar cálculos flexibles y eficientes de Wiki3DRank. Los usuarios pueden explorar los valores cambiantes de las métricas cuando se incluyen en los cálculos características opcionales como el número de enlaces entrantes y salientes a los artículos o las relaciones entrantes y salientes.

5. Discusión

Los resultados presentados son bastante explícitos en su planteamiento, tanto desde el punto de vista de la ejecución como del cálculo. Se ha buscado en todo momento mantener la sencillez en el proceso, de forma que sea fácilmente replicable y observable. Esta sección se centra en la discusión en tres aspectos de muy diferente naturaleza, pero que parecen relevantes para la comunidad investigadora y profesional: a) la eficiencia de ejecución de los cálculos; b) la incorporación de más componentes a *Wiki3DRank*; c) el refinamiento de los resultados mediante propiedades de dominio.

5.1. Eficiencia en la obtención de datos

Uno de los aspectos más relevantes a debatir sobre el método de cálculo es la eficiencia de la obtención del indicador N_{Words}. En este trabajo se ha mostrado cómo dicho indicador se calcula a partir de la suma del número de palabras de cada artículo relacionado con el ítem en cuestión y obtenidos a través de XTools. El principal inconveniente de este método es la necesidad de realizar una conexión a la API de XTools por cada artículo en cada Wikipedia, dado que el uso de conexiones asíncronas está limitada por el servidor de XTools.

La experiencia con la herramienta *Wiki3DRank Calculation* ha mostrado que el número óptimo de conexiones concurrentes es de 35. Esto significa que para aquellos ítems con un elevado número de artículos se precisan varias conexiones asíncronas. Esto provoca cierto nivel de demora en la obtención de los datos, y rebaja la calidad de la experiencia de usuario. Una alternativa práctica es calcular N_{Words} para cada ítem limitando a 35 el número de artículos de Wikipedia de manera que puedan recuperarse todos los datos en una única conexión. La versión rápida de Wiki3DRank Calculation adopta este método, pero seleccionando siempre las 35 wikipedias con mayor número de artículos en los que el ítem tiene alguna equivalencia. En este caso al indicador se le ha denominado N_{Words_fast} y al valor del ranking $Wiki3DRank_{fast}$.

Otra aproximación diferente, que podría ser interesante, es el uso de una medida alternativa al número de palabras de cada uno de los artículos. Dicha alternativa consideraría el valor del esfuerzo de edición a nivel global de la edición de Wikipedia del artículo, en vez de considerar los artículos de forma individual. De esta forma se sumaría la media de palabras de todos los artículos de cada edición de Wikipedia. Por lo tanto, este indicador, al que se ha denominado N_{Words_wm}, se calcularía sumando la media de palabras por artículo de cada una de las Wikipedias en las que tiene un artículo equivalente el ítem de Wikidata. El número total de artículos y de palabras se obtiene de la página de estadísticas de cada Wikipedia, cuyos datos se almacenan en el servidor y se pueden actualizar periódicamente, a través de un script que los almacena como fichero JSON. El cálculo de palabras por artículo de cada Wikipedia no varía significativamente, por lo que los datos se obtendrían a partir de la carga de un fichero estático que podría actualizarse periódicamente. Al valor del ranking calculado con este método se le ha llamado $Wiki3DRank_{wm}$.

A partir del análisis de casos se ha realizado un estudio comparando los resultados del ranking original que utiliza N_{Words} con los resultados en los que se calcula Wiki3DRank con N_{Words_wm} y N_{Words_fast}. Los datos de la Figura 5 muestran la coincidencia del ranking en función del tamaño de la muestra de los ítems ordenados inversamente por el valor de Wiki3DRank. Puede verse como en los tres conjuntos de datos seleccionados (ríos, películas y obras literarias) los valores que se obtienen con $Wiki3DRank_{fast}$ son mucho más precisos que los obtenidos con $Wiki3DRank_{wm}$. Por lo tanto, se podría concluir que utilizando únicamente las treinta y cinco ediciones de Wikipedia con mayor número de artículos se obtienen valores muy cercanos al Wiki3DRank original al tiempo que se consigue una mayor eficiencia en la obtención de datos para el cálculo al poder recuperar de XTools todos los datos de las palabras de los artículos de las ediciones de Wikipedia para calcular N_{Words_fast} con una única conexión.

Sin embargo, aunque parezca que existen diferencias significativas, es necesario un estudio más profundo, ya que el número de obras comunes en las listas obtenidas con ambos métodos es mayor cuando se aumenta el número de ítems. En ambas litas, de los 150 primeros elementos, se observa que se comparten el 80,6% de las obras, con 500 elementos, comparten el 80,8%, y con 1000 elementos, ambas listas comparten el 78,8% de las obras.

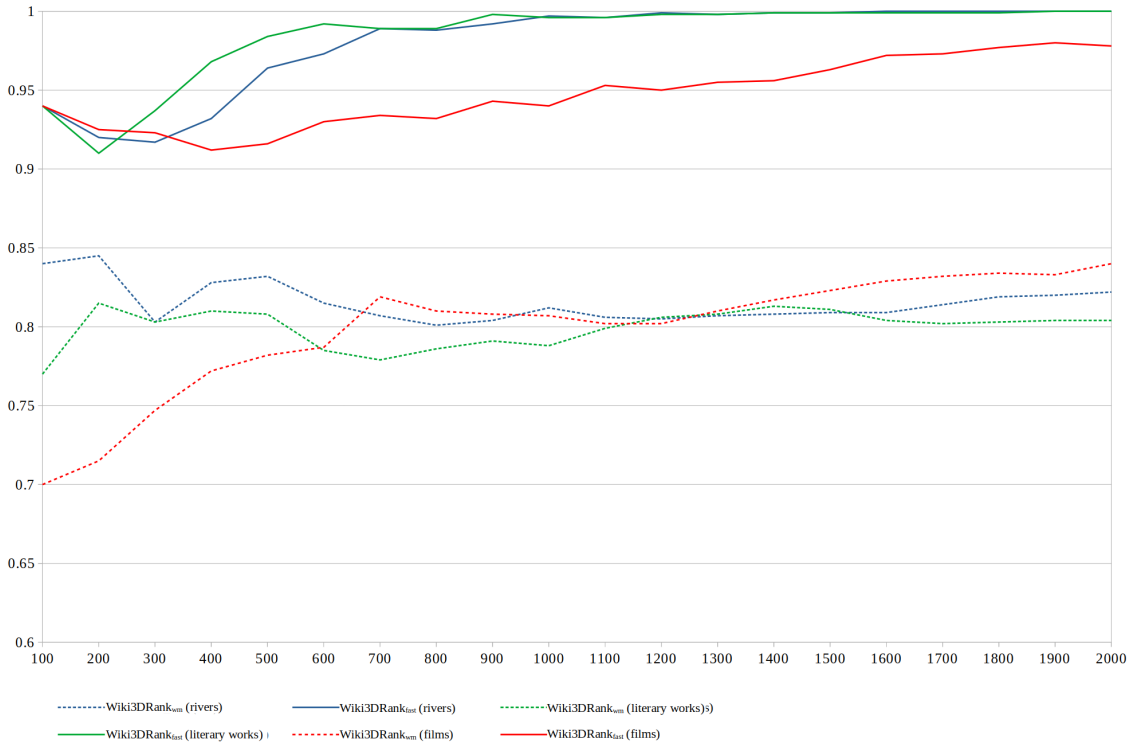

Figura 5. Grado de coincidencia (eje Y) en las N primeras posiciones del Ranking (eje X) del Wiki3DRank comparado con Wiki3DRank$_{fast}$ y Wiki3dRank$_{wm}$ (Datos: enero 2024)

5.2. Aumento del número de componentes: de 3D a 360º

Otro aspecto que resulta sugerente es la incorporación de componentes adicionales al vector. Se trataría de una ampliación del modelo que tendrían en cuenta los siguientes indicadores además de los ya mencionados anteriormente. La siguiente tabla muestra una descripción completa de los mismos:

Indicador	*Descripción*	*Fuente*	*Método*
N_{Uprops}	Frecuencia de uso de propiedades diferentes en un mismo ítem, que no sean del tipo identificadores externos.	Wikidata	SPARQL WDQS
$N_{Inprops}$	Número de relaciones entrantes a un ítem.	Wikidata	SPARQL WDQS
$N_{Uinprops}$	Número de relaciones entrantes desde diferentes ítems.	Wikidata	SPARQL WDQS
$N_{Idprops}$	Número de afirmaciones con propiedades de identificadores externos en un mismo ítem.	Wikidata	SPARQL WDQS
$N_{Uidprops}$	Frecuencia de uso de propiedades diferentes de identificadores externos en un mismo ítem.	Wikidata	SPARQL WDQS
$N_{Section}$	Número total de secciones en todos los artículos de un ítem	Wikipedia	XTools Prose
N_{Refs}	Número total de referencias en todos los artículos de un mismo ítem.	Wikipedia	XTools Prose
N_{Urefs}	Número total de referencias únicas en todos los artículos de un mismo ítem.	Wikipedia	XTools Prose
N_{Lext}	Número total de enlaces externos en todos los artículos de un mismo ítem	Wikipedia	XTools Links
N_{Lout}	Número total de enlaces salientes a otros artículos en todos los artículos de un mismo ítem.	Wikipedia	XTools Links
N_{Lin}	Número total de enlaces entrantes desde otros artículos hacia los diferentes artículos de un mismo ítem	Wikipedia	XTools Links

Table V. Aditional indicators for Wiki3DRank calculation

En la aplicación presentada, Wiki3DRank Calculation, se ha incluido como elemento opcional por el usuario el uso de todas estas variables para calcular la medida de relevancia. Esto permite el análisis y evaluación de los resultados obtenidos sobre pequeñas muestras. Para demostrar su validez genérica, se hace necesario desarrollar una

investigación que no entra en el alcance de esta publicación.

A este respecto en lugar de incorporar todos los elementos disponibles para la obtención de una métrica única, parece una vía más eficiente alcanzar resultados análogos en la práctica a través de los datos justos mínimos, facilitando su interpretación y validación. En las disciplinas relacionadas con informetría, más no siempre es mejor, ni proporciona más claridad para evaluar o conocer recursos de información (Torres-Salinas; Robinson-García; Jiménez-Contreras, 2023). A partir de cierto punto, los incrementos marginales previsibles al incorporar más variables parece no suponer una mejora de calidad apreciable, y sí una complicación para la explicación y comprensión de la medida propuesta.

5.3. Refinamiento mediante propiedades de dominio: el caso de las obras creativas

Resulta difícil disponer de una medida universal que satisfaga las características de todos los casos. La medición de la relevancia a través de los datos de Wikidata-Wikipedia es más significativa en determinados dominios. En la sección anterior se mostraron dos ejemplos relacionados con las obras creativas, que son objetos culturales u objetos de conocimiento que disponen de una amplia colección de instrumentos de catalogación, compilación, difusión y evaluación. En este ámbito, concretamente en las obras cinematográficas y literarias, los resultados muestran un cierto «presentismo» que parece favorecer a las obras recientes. Este fenómeno podría explicarse por la mayor atención prestada por la comunidad editora a obras de reciente impacto social (como el estreno de grandes películas, bestsellers, campañas promocionales, etc.). Para tener en cuenta este efecto, puede ser interesante añadir un componente adicional que represente la antigüedad de las obras.

Este nuevo componente daría mayor peso a las obras con un historial de publicación o creación más largo. Es importante destacar que, mientras que los componentes anteriormente mencionados se incorporan directamente al cálculo del vector, en este caso se trata de definir y justificar cómo se integrará en el cálculo.

El primer paso es identificar las propiedades de dominio candidatas y analizar su implementación, especialmente las relaciones de subpropiedades RDFS, y su uso por parte de la comunidad para el tipo de objetos del dominio. En el caso de las obras literarias, investigaciones previas indican que el «presentismo» está bien equilibrado en la métrica (Pastor-Sánchez; Saorín; Baños-Moreno, 2023). Sin embargo, en el ámbito audiovisual, que se rige por criterios más marcados de consumo rápido y masivo, sí parece distorsionar el acercamiento a resultados más consensuados sobre la evaluación de las obras.

Las propiedades que se usarian para obtener la antigüedad de la obra son P577 (fecha de publicación) y, alternativamente, P571 (fecha de creación). Estas propiedades reflejan la instanciación de una obra para su difusión, que es un elemento fundamental en cualquier esquema de metadatos, como Dublin Core o Schema.org. Las propiedades de fecha se organizan mediante un esquema relativo de relaciones de subclase y subpropiedad, pero su lógica no es muy rigurosa.

Este trabajo propone en este caso calcular la diferencia entre la fecha actual y la fecha de la obra. El resultado se utilizaría para obtener un nuevo componente N_{Date} y se incorporaría al vector de cálculo de *Wiki3DRankDate*. De este modo, la ecuación de cálculo original podría ser la siguiente:

$$a = log\ (1 + N_{Wikis})$$

$$b = log\ (1 + N_{Props})$$

$$c = log\ (1 + N_{Words})$$

$$d = log(1 + N_{Date})\ ;\ N_{Date} = Year_{current} - Year_{pub}$$

$$Wiki3DRank = \sqrt{a^2 + b^2 + c^2 + d^2}$$

Un ejemplo de aplicación de *Wiki3DRank*$_{Date}$ puede verse en la Tabla VI, en la página siguiente, en donde puede verse como aparecen en las primeras veinte posiciones algunos ítems de películas que podrían considerarse clásicos del cine y que con el cálculo original de *Wiki3DRank* estaban relegados a posiciones más bajas del ranking.

Esta aproximación preliminar sugiere que los resultados mantienen un fuerte componente de actualidad, por no hablar de un sesgo netamente americano que apenas capta la globalidad del cine como medio de expresión artística, más allá de una industria del entretenimiento. Es importante resaltar que en algunos casos es necesario poner un límite a N_{Date}. Es posible encontrar en el dominio de obras creativas o eventos fechas muy lejanas. Mientras que el cine tiene una historia de poco más de un siglo, la literatura, la pintura y otras artes tienen una larguísima tradición, habría que tener en cuenta el desequilibrio de los clásicos milenarios o centenarios. En estos casos el uso sin límite de N_{Date} puede introducir el efecto contrario en los resultados del ranking. Es decir, podría darse el caso de encontrar en las primeras posiciones ítems vinculados a fechas lejanas puesto que se ponderarían excesivamente al

alza. Por este motivo se precisaría un estudio más detenido y específico para construir correctamente este modulador, valorando el porcentaje máximo de aportación de N_{Date} para el cálculo de *Wiki3DRank*.

Item	*Label*	*Wiki3DRank*	*Item*	*Label*	*Wiki3DRank$_{Date}$*
Q44578	Titanic	14.12210	Q44578	Titanic	14.26255
Q24871	Avatar	13.79666	Q17738	Star Wars: Episode IV – A New Hope	14.01840
Q17738	Star Wars: Episode IV – A New Hope	13.76991	Q2875	Gone with the Wind	14.00175
Q163872	The Dark Knight	13.73499	Q134430	Snow White and the Seven Dwarfs	13.92978
Q47703	The Godfather	13.70764	Q47703	The Godfather	13.90140
Q104123	Pulp Fiction	13.66943	Q132689	Casablanca	13.82313
Q23781155	Avengers: Endgame	13.65912	Q103474	2001: A Space Odyssey	13.71408
Q102438	Harry Potter and the Philosopher's Stone	13.61317	Q102438	Harry Potter and the Philosopher's Stone	13.66237
Q2875	Gone with the Wind	13.61120	Q103569	Alien	13.62753
Q134430	Snow White and the Seven Dwarfs	13.56524	Q184843	Blade Runner	13.61312
Q23780914	Avengers: Infinity War	13.46630	Q104123	Pulp Fiction	13.59927
Q182218	The Avengers	13.44259	Q24871	Avatar	13.56592
Q91540	Back to the Future	13.42724	Q41483	The Good, the Bad and the Ugly	13.55800
Q103474	2001: A Space Odyssey	13.42255	Q91540	Back to the Future	13.55078
Q18407657	Captain America: Civil War	13.41430	Q23781155	Avengers: Endgame	13.52241
Q23780734	Black Panther	13.36900	Q24815	Citizen Kane	13.52018
Q14171368	Avengers: Age of Ultron	13.36220	Q180098	Ben-Hur	13.51389
Q134773	Forrest Gump	13.35778	Q483941	Schindler's List	13.49594
Q132689	Casablanca	13.33334	Q42051	Star Wars: Episode III – Revenge of the Sith	13.43629
Q483941	Schindler's List	13.32423	Q134773	Forrest Gump	13.43046

Table VI. Comparativa de los primeros veinte primeros resultados entre Wiki3DRank y Wiki3DRank$_{Date}$ del conjunto de datos de películas (Datos: enero 2024)

6. Conclusiones and trabajo futuro

A lo largo de este trabajo se ha presentado una metodología de cálculo de *Wiki3DRank* no solo sencilla en su formulación a partir del uso de un espacio vectorial, sino también de fácil lectura e interpretación sencilla. Aplicado en varios casos de uso sobre objetos de conocimiento, se aprecia que su funcionamiento sobre objetos que poseen cierto "impacto social" (difusión, audiencia, valor, población, territorio, repercusiones sobre la sociedad, etc) refleja notabilidad o relevancia. Medir tiene un valor muy visible en el negocio social de la atención, pero a menudo se obvia que también puede tenerlo en los sistemas de organización del conocimiento. Los conceptos de una taxonomía o los elementos de un listado de autoridades no tienen todos la misma importancia. La existencia de herramientas de fácil acceso, procedimientos transparentes y reproducibles, así como métricas estandarizadas u ofrecidas por proveedores de confianza, para medir diferentes aspectos de la sociedad del conocimiento ayuda al desarrollo del lenguaje de las humanidades digitales y el campo de la información.

En cuanto a los tres elementos *core* usados para construir la métrica, N_{Wikis}, N_{Props} y N_{Words}, sería necesario analizar con mayor detalle las correlaciones entre las variables, para comprender mejor su aportación en la construcción del valor de ranking y para ejecutar procesos de clusterización. También parece conveniente explorar y validar la posibilidad del uso de conexiones o enlaces entrantes y/o salientes, como otra capa para entender la calidad del contenido presente en los artículos, aunque aquí se plantea aquí la dificultad operativa para conocer los valores generales o relativos de cada nodo en un grafo. El enfoque de reducción de la dimensionalidad, seleccionando variables que, combinadas, permiten filtrar, agrupar y rankear, es viable, aunque parece sensato considerar que para obtener resultados más afinados hay que manejar también variables de dominio, que sean significativas para un tipo de objeto de conocimiento específico. El uso de variables específicas para cada ámbito, como la

fecha en las obras creativas, los premios a los artistas, e incluso su ponderación en el cálculo de la puntuación final, implica una rigurosa construcción y validación de indicadores compuestos (Blasco-Blasco, Rodríguez-Castro; Tuñez-López, 2020). Esta podría ser, sin duda, una interesante vía de investigación futura que implique la integración de datos ajenos al ámbito Wikimedia. El uso del número de palabras como variable para reflejar la profundidad del contenido supone una limitación de lo que supone el valor de una información: la estructuración del contenido en apartados, la existencia de notas, bibliografía o ilustraciones reflejarían de forma más rica la calidad de un contenido enciclopédico. También es necesario tener en cuenta que existe una contraprestación entre riqueza de los datos y velocidad de proceso cuando se usa Xtools en lugar de las APIs de MediaWiki.

Para delimitar el dominio de análisis, se constata que hay dificultades para el uso de la taxonomía de clases de Wikidata para seleccionar con precisión y exhaustividad recursos de un mismo tipo con el fin de hacer análisis de elementos afines de un mismo dominio o campo. Este es un problema diferente al del mero cálculo, pero que complica su aplicabilidad para estudios sectoriales porque la medida de ranking adquiere sentido en un conjunto de elementos comparables.

Las metodologías de análisis de redes no son las únicas aplicables sobre grandes conjuntos de datos conectados. En muchos casos hay aproximaciones menos costosas, que reducen las barreras de acceso o que proporcionan datos interpretables con claridad suficiente. Es posible obtener resultados sin procesar y reprocesar un dataset completo. Esta estrategia facilita la asignación de valoraciones métricas y su actualización, dado que el valor no depende del estado de todos los elementos, sino que se deriva del cálculo de propiedades individuales previamente validadas y consensuadas. La obtención de valores *on-the-fly* puede ser útil para enriquecer otros sistemas de descubrimiento de información, en el proceso de recomendación y filtrado de resultados, e incluso una forma de reproducir medidas aproximadas de centralidad sin tener acceso al grafo completo.

Otra limitación que hay que considerar es que aquellos objetos de conocimiento del ámbito de la cultural local, que suelen aparecen en una única Wikipedia, están penalizados en Wiki3DRank. El modelo utilizado da por supuesto que el conocimiento es universal y está reflejado en el mayor número de idiomas posible. Los objetos globales, por su propia naturaleza objetiva, tamaño, e impacto tienen ventaja obvia sobre aspectos vinculados a una cultura o región.

Parece oportuno mencionar que la enorme atención que acapara el término Big Data hace olvidar enfoques muy útiles, basados en conjuntos de datos grandes (Long data) y en el uso eficiente de pocas variables representativas. La economía de medios contiene su propia propuesta de valor, para adquirir *insights* sobre sus objetos de estudio, frente al despliegue de arsenales analíticos y de datos de gigantes de la red. Los datos de tamaño oceánico – internet - a menudo actúan en sustitución de datos no disponibles sobre los fenómenos que queremos observar. No siempre hay datos, o son muy imperfectos, pese a la alucinación colectiva que cree en lo contrario de forma automática (Borgman, 2017). El mandato de “lo inteligente” ha de ser entendido con sutileza, como la necesidad de la dimensión suficiente, y de una suficiente comprensión de la variedad de registros interpretativos (Halpern; Mitchell, 2022).

La definición acumulativa que aplica Wiki3DRank supone en la práctica que un elemento no puede reducir su valor con el tiempo (a excepción de que los artículos fueran eliminados o condensados, lo que es poco habitual) y cabría realizar simulaciones sobre el costo de nuevos elementos para competir con los ya establecidos o la desnaturalización del ranking en el caso de muchos elementos con valores muy similares.

También hay que tener en cuenta que el conocimiento de un sistema de medición puede facilitar su manipulación. La cuantificación del ranking implica que pueden realizarse acciones para generar el tipo de datos que aumentan su ranking. Por no examinar acciones de naturaleza destructiva, en lo que respecta a una “sobrealimentación” de un recurso para mejorar su ranking, nuestra somera estimación superficial sugiere que las acciones de edición masiva en muchos idiomas son costosas de orquestar, y que el enriquecimiento de registros creando nuevas declaraciones tiene un efecto muy controlado. Sin embargo, se aprecian dos excepciones que pueden adquirir dimensiones más acusadas: por un lado se ha detectado que el uso puntual de Wikidata como catálogo exhaustivo de las ediciones de una obra, puede generar un volumen de datos que impacte en el ranking (Situación detectada en el ítem “La guerra de los mundos” de George Orwell, con 6000 declaraciones de la propiedad P655 [has edition]). En segundo lugar, la disponibilidad masiva de motores de inteligencia artificial generativa con capacidad multilingüe, puede simplificar bombardeos selectivos de cantidades de texto sobre muchas ediciones de Wikipedia, que pueden afectar también al ranking. La primera situación puede tenerse en cuenta conociendo la actividad del grupo de interés en

avanzar hacia que Wikidata sea una base de datos bibliográfica y catalográfica global; la segunda situación abre un panorama mucho menos fácil de delimitar y observar.

Finalmente destacar la claridad y sencillez del cálculo vectorial propuesto, al permite añadir componentes que se incorporan a la mecánica de cálculo y sin necesidad de construir alambicadas métricas compuestas.

Es voluntad de los autores del trabajo poner a disposición pública, además del código fuente de los datos y los scripts usados en el estudio, mejoras continuas en la aplicación web presentada, que permita explorar el Wiki3DRank tanto de colecciones amplias de ítems (tipos, clases) como de selecciones de itemas ad-hoc y compararlos de forma ágil.

Acceso abierto a los datos y scripts de investigación

Dataset and data processing scripts: https://doi.org/10.5281/zenodo.10576041

Source code for the Wiki3dRank Calculation web application: https://github.com/j-pastor/wiki3drank

Wiki3dRank Calculation web app: https://gicd.inf.um.es/wiki3drank

Referencias

Ahnert, Ruth; Ahnert, Sebastian; Coleman, Catherine; Weingart, Scott (2020). The Network Turn: Changing Perspectives in the Humanities. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108866804

Anderson, Chris (2014). The Longer Tail Why the Future of Business is Selling Less of More. New York: Hachette Books.

Beytía, Pablo; Schobin, Janosch (2020) Networked Pantheon: a Relational Database of Globally Famous People. // Research Data Journal for the Humanities and Social Sciences. 5, 50-65. https://doi.org/10.1163/24523666-00501002

Bianchini, Carlo; y Sardo, Lucia (2022). Wikidata : a new perspective towards universal bibliographic control. // JLIS. 13:1, 291-311. https://doi.org/10.4403/jlis.it-12725

Blank, Grant (2007). Critics, Ratings, and Society. Lanham: Rowman and Littlefield.

Blasco-Blasco, Olga; Rodríguez-Castro, Marta; Túñez-López, Miguel (2020). Composite indicators as an innovative methodology for Communication Sciences: implementation for the assessment of European public service media". // Profesional de la información. 29, n. 4, e290437, 2020. https://doi.org/10.3145/epi.2020.jul.37

Borgman, Christine L. (2017). Big data, little data, no data. Cambridge, Massachusetts: The MIT Press. https://doi.org/10.7551/mitpress/9963.001.0001

Brown, Andrew (2011). A brief history of encyclopaedias: from Pliny to Wikipedia. Londres: Hesperus.

Halpern, Orit; Mitchell, Robert (2022) The smartness mandate. Cambridge, Massachusetts: The MIT Press. https://doi.org/10.7551/mitpress/14623.001.0001

Hill, Benjamin Mako; Shaw, Aaron (2020). The Most Important Laboratory for Social Scientific and Computing Research in History. // Reagle, Joseph; Koerner, Jackie (eds.). Wikipedia @ 20: Stories of an Incomplete Revolution. Cambridge, Massachusetts: The MIT Press. https://doi.org/10.7551/mitpress/12366.001.0001

Lewoniewski, Włodzimierz; Węcel, Krzysztof; Abramowicz, Witold (2019). Multilingual Ranking of Wikipedia Articles with Quality and Popularity Assessment in Different Topics. // Computers. 8:3, 60. https://doi.org/10.3390/computers8030060

McDowell, Zachary J.; Vetter, Matthew A (2022). Wikipedia and the Representation of Reality. New York: Routledge. https://doi.org/10.4324/9781003094081

Minguillón, Julia; Lerga, Maura; Aibar, Eduard; Lladós-Masllorens, Josep; y Meseguer-Artola, Antoni (2017). Semi-automatic generation of a corpus of Wikipedia articles on science and technology. // El Profesional de la Información. 26:5, 995-1004. https://doi.org/10.3145/epi.2017.sep.20

Miquel-Ribé, Marc (2019). The Sum of Human Knowledge? Not in One Wikipedia Language Edition. Wikipedia@20. https://wikipedia20.mitpress.mit.edu/pub/26ke5md7/release/15

Moás, Pedro Miguel; Teixeira Lopes, Carla (2023). Automatic Quality Assessment of Wikipedia Articles: A Systematic Literature Review. // ACM Computing Surveys. 56:4, article 95. https://doi.org/10.1145/3625286

Nielsen, Finn Årup (2012). Wikipedia Research and Tools: Review and Comments. http://doi.org/10.2139/ssrn.2129874

Piscopo, Alessandro; y Simperl, Elena (2018). Who Models the World?: Collaborative Ontology Creation and User Roles in Wikidata. // Proceedings of the ACM on Human-Computer Interaction. 2:CSCW, Article 141. https://doi.org/10.1145/3274410

Reznik, Ilia; Shatalov, Vladimir (2016). Hidden revolution of human priorities: An analysis of biographical data from Wikipedia. // Journal of Informetrics. 10:1, 124-131. https://doi.org/10.1016/j.joi.2015.12.002

Shenoy, Kartik; Ilievski, Filip; Garijo, Daniel; Schwabe, Daniel; Szekely, Pedro (2022). A study of the quality of Wikidata. Journal of Web Semantics. 72, 100679. https://doi.org/10.1016/j.websem.2021.100679

Skiena, Steven; Ward, Charles B. (2014). Who's bigger? Where historical figures really rank. Cambridge: Cambridge University Press.

Torres-Salinas, Daniel; Robinson-García, Nicolás; Jiménez-Contreras, Evaristo (2023). The bibliometric journey towards technological and social change: A review of current challenges and issues. // Profesional de la información. 32:2, e320228. https://doi.org/10.3145/epi.2023.mar.28

Este artículo es la versión en español
del artículo anterior publicado en el mismo número

Enviado: 2024-03-14. Segunda versión: 2024-05-21.
Aceptado: 2024-05-23.