

Título del Proyecto:

Sistema Recomendador de libros.

Dominio de aplicación: Recomendación de libros y autores.

Estudiantes: Lidia González Martín y Juan Broto Ortega.

1. Descripción del problema

Los usuarios buscan libros basándose en los géneros y autores de libros que han leído previamente y que les han gustado, pero la información acerca de los libros es muy amplia y dispersa, pudiendo agobiar a los usuarios que quieren una nueva lectura.

Este proyecto desarrolla un sistema recomendador enriquecido con ontologías que comprende relaciones semánticas de géneros. Permite razonar por similitud sobre los géneros. Así, el sistema puede recomendar obras similares a los usuarios basándose en sus estilos de escritura favoritos, teniendo en cuenta elementos tan relevantes como el autor del libro o la editorial.

Se han identificado diversos problemas que justifican la necesidad de un recomendador de libros. En primer lugar, existe una sobrecarga informativa, los catálogos digitales contienen millones de títulos y las recomendaciones están basadas en la popularidad, sin tener en cuenta relaciones entre géneros.

2. Fuentes de conocimiento/datos analizadas

Se ha realizado el análisis de diversas fuentes de datos bibliográficos y ontologías existentes para construir un sistema de recomendación robusto y semánticamente rico. El objetivo era integrar fuentes que cubrieran dos necesidades: proporcionar estructura taxonómica completa y aportar instancias actualizadas de libros. Con respecto a la estructura taxonómica, se tuvieron en cuenta diversas fuentes:

- **Wikidata.** Empleamos wikidata como fuente principal para la extracción de la jerarquía de géneros literarios narrativos. Mediante consultas SPARQL, extrajimos la jerarquía completa de géneros narrativos partiendo de la entidad "género narrativo" (Q1318295). La relación "subclase de" (P279) permitió construir un árbol taxonómico con tres niveles de profundidad, capturando relaciones como "novela policíaca" → "ficción detectivesca" → "género narrativo".
- **Otras ontologías empleadas.** Con la finalidad de enriquecer la jerarquía obtenida de wikidata mapeamos otras ontologías externas, este es el caso de **BIBO** (Bibliographic Ontology), que aporta un estándar académico para metadatos bibliográficos, **Schema.ORG**, que aporta relaciones para las propiedades de autor, género e isbn, **Dublin Core** que aporta relaciones para las propiedades de creador, editorial y fecha de lanzamiento y **FOAF**, útil para las relaciones sociales, en este caso para autor.

Fuente evaluadas

Para la obtención de libros se evaluaron varias fuentes de datos. En primer lugar, se probó **Google Books API** que con su url permite acceder de forma estructurada a múltiples datos bibliográficos, como títulos y autores, información editorial consistente, ISBN validados, fechas de publicación y descripciones y clasificaciones de edad. Google Books resultó ser una fuente muy completa con tiempo de respuesta estables.

Por otro lado, se analizó también **OpenLibrary** en la que se encontraron algunas limitaciones técnicas. Entre ellas la inestabilidad en la información de cada libro, con campos variables e información frecuentemente incompleta, así como problemas recurrentes en los identificadores, en particular ISBN ausentes, lo mismo ocurría con la información de las editoriales.

El tiempo de las consultas también resultó un problema, la arquitectura de los end-points de OpenLibrary obliga a la realización de múltiples consultas para obtener toda la información necesaria y cada consulta tiene tiempos de respuesta prolongados. Estas limitaciones no invalidan el valor de OpenLibrary como repositorio amplio y abierto, pero sí limitan su uso, por ello decidimos no utilizarla para la creación .

DBpedia también fue objeto de evaluación como posible fuente adicional, mediante un método específico que aprovechaba su interconexión con Wikidata a través de relaciones `*owl:sameAs*`. Sin embargo, en el ámbito concreto de los libros, especialmente los contemporáneos, se detectó una cobertura irregular y la ausencia frecuente de metadatos esenciales, como ISBN o fechas de publicación precisas. Esta falta de completitud redujo su utilidad práctica para los objetivos planteados, aunque se mantiene como una fuente relevante especialmente por su relación con Wikidata.

Finalmente, se utilizó **Wikidata** para enriquecer la ontología. Aprovechando que la jerarquía se extrajo de la misma fuente y que se conservaron sus identificadores, Wikidata se convirtió en una fuente especialmente interesante para completar con más libros. Los datos se extrajeron mediante consultas SPARQL, durante cuyo proceso surgieron diversos problemas. El primero de ellos fue la limitación en la frecuencia de las consultas, que resultó ser extremadamente baja, provocando que muchas de ellas no se procesaran debido a excepciones del servidor. A esto se sumó el elevado tiempo de latencia en la ejecución de las consultas, lo que prolongó considerablemente la generación del grafo de conocimiento. Otro de los problemas detectados fue la falta de ciertos campos, como Wikidata es un grafo de conocimiento de propósito general y no específico dentro del dominio de los libros, en algunos casos las descripciones, el ISBN, la editorial u otros campos específicos se retornaron vacíos.

Estrategia de integración

Cada fuente se procesó de forma independiente. En primer lugar, Wikidata proporcionó la estructura taxonómica fundamental. Por su parte las ontologías externas (BIBO, Schema.org, Dublin Core, FOAF) establecieron puentes semánticos con estándares ampliamente adoptados. Y finalmente Google Books API, DBpedia y WikiData enriquecieron la ontología con instancias actualizadas de libros.

La integración se realizó mediante mapeos semánticos, estableciendo equivalencias entre propiedades y clases de diferentes vocabularios. Por ejemplo, la propiedad `onto:tieneAutor` se declaró equivalente a `schema:author`, `dc:creator` y `bibo:authorList`, creando una red de relaciones interoperable.

Parte de las instancias de libros se obtuvieron de Google Books mediante consultas URL basadas en el label del género obtenido de Wikidata. No obstante, esta relación no resultó completamente estable, por lo que algunos géneros de Wikidata quedaron sin instancias asociadas. Debido a ello, se incorporaron DBpedia y la propia Wikidata como fuentes externas adicionales.

3. Justificación del enfoque basado en conocimiento

- ¿Qué pueden aportar las ontologías frente a un enfoque basado únicamente en datos?

En un sistema de recomendación, las ontologías ayudan a superar las limitaciones de los enfoques que se basan solo en datos numéricos, porque añaden significado y reglas claras sobre cómo se relacionan los conceptos. Mientras que los modelos basados en datos se limitan a detectar patrones y coincidencias en los datos, las ontologías permiten describir de forma explícita las relaciones que explican esos patrones.

Las ontologías permiten representar el conocimiento del dominio de manera estructurada. En el caso de los libros, hacen posible definir claramente la jerarquía de los géneros literarios y las relaciones entre autores, obras y géneros. Gracias a esto, el sistema puede razonar sobre las preferencias de un usuario. Por ejemplo, si una persona muestra interés por la “novela negra” y este género forma parte de la “ficción detectivesca”, el sistema puede inferir que también podrían resultarle atractivos otros géneros cercanos, como el thriller policíaco.

Otro aspecto importante es la integración de datos procedentes de distintas fuentes. Mediante relaciones semánticas estándar, es posible vincular información de orígenes diferentes, como Wikidata, Google Books o DBpedia, aunque usen estructuras o nombres distintos. Esto facilita combinar y reutilizar datos que, con procesos tradicionales, resultarían difíciles de unificar.

Además, las ontologías mejoran la explicabilidad de las recomendaciones. Al basarse en un modelo comprensible, el sistema puede justificar por qué sugiere un libro, por ejemplo, indicando que pertenece a un género relacionado con otros que el usuario ya ha leído o que comparte autor con obras previas. Esto contrasta con muchos modelos de aprendizaje automático, cuyos resultados son difíciles de interpretar.

- Indica los enunciados de algunas preguntas o consultas que van a ser útiles y que permiten validar el conocimiento

¿Qué libros son similares a X basándose en su género?

¿Teniendo en cuenta a los demás usuarios con gustos similares qué libros le podría gustar a un usuario A sabiendo que le gustan los libros 1, 2 y 3?

4. Dificultades encontradas

Además de las dificultades mencionadas anteriormente relacionadas con las diferentes fuentes de datos analizadas (ver Sección 2) nos surgieron las siguientes dificultades durante la implementación del sistema:

- **Similitud Item-To-Item.** Al inicio nuestra taxonomía de géneros estaba estructurada de tal manera que los libros eran RDF.type de los géneros, pero esto resultó ser problemático ya que al navegar por los diferentes géneros de la jerarquía el algoritmo no conseguía llegar a los mismos. Por ello tuvimos que redefinir la estructura para que los libros pasasen a ser RDF.type de una clase denominada Librosxxi y establecer una propiedad con dominio Librosxxi y rango géneros llamada tieneGenero.
- **Filtrado colaborativo.** Los usuarios se relacionan con los libros a través de una propiedad llamada leGusta que tiene como dominio a los propios usuarios y de rango a la clase Librosxxi. Inicialmente calculábamos la similitud entre dos perfiles a través de la métrica de Jaccard. Sin embargo, ante un volumen de datos medio y perfiles con baja superposición, esta métrica no permitía obtener valores de similitud suficientemente elevados entre usuarios, lo que dificultaba la generación de recomendaciones. Por ello decidimos optar por una adaptación de la métrica en la que se ponderó el denominador de forma que la fórmula únicamente tuviera en cuenta el perfil con menor número de libros.
- **Falta de Serendipia.** Una de las dificultades que encontramos fue la falta de serendipia en las recomendaciones. Esta falta de diversidad vino originada por la forma en la que calculamos la similitud Item to Item. Para resolverlo se decidió agregar pesos de forma que se tuviera en cuenta autores o editoriales en común, dando más peso al autor (0.5) que a la editorial (0.2). Se agregó a su vez un factor de aleatoriedad que suma aleatoriamente un factor de hasta 0.1.

5. Conclusiones

En este trabajo se ha desarrollado un sistema recomendador de libros basado en conocimiento, con el objetivo de reducir la sobrecarga informativa presente en los catálogos digitales actuales. El uso de ontologías ha permitido crear jerarquía de géneros literarios y las relaciones entre libros, autores y editoriales, posibilitando un razonamiento semántico que va más allá de las recomendaciones basadas en popularidad o coincidencias directas.

La integración de múltiples fuentes de datos, Wikidata, Google Books API, y DBPedia junto con ontologías estándar como BIBO, Schema.org, Dublin Core y FOAF, ha dado lugar a un modelo interoperable y escalable. Este enfoque ha facilitado la generación de recomendaciones más coherentes y explicables.

Asimismo, la combinación de similitud basada en ítems y filtrado colaborativo, junto con ajustes en las métricas de similitud y la introducción de factores de diversidad, ha permitido mejorar la calidad de las recomendaciones, abordando problemas como la baja superposición entre perfiles y la falta de serendipia. En conjunto, los resultados obtenidos confirman la idoneidad de los enfoques basados en conocimiento para el dominio de la recomendación de libros, habiendo alcanzado con nuestra implementación los objetivos planteados inicialmente.

6. Trabajo Futuro

Como trabajo futuro, se propone ampliar y refinar la ontología incorporando nuevos géneros y relaciones semánticas, así como mejorar los mecanismos de adquisición y actualización de datos para reducir problemas de latencia y falta de completitud.

También sería de interés explorar enfoques híbridos que combinen el razonamiento ontológico con técnicas de aprendizaje automático, con el fin de capturar similitudes no explícitas y mejorar la diversidad de las recomendaciones. Finalmente, la evaluación del sistema con usuarios reales permitiría comprobar su utilidad y precisión.